DEPARTMENT OF STATISTICS

The Ohio State University

OSU

COLUMBUS, OHIO

Recent Applications of Variational Techniques in Statistics

J. S. Rustagi*

Department of Statistics
The Ohio State University
Columbus, Ohio 43210

Technical Report No. 186

August, 1979

* Supported by Contract NOOO 14-78-C-0543 (NR 042-403) by Office of Naval Research. Reproduction in whole or in part is permitted for any purpose of the United States Government.

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
4

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
6) Recent Applications of Variational Techniques in Statistics

5. TYPE OF REPORT & PERIOD COVERED
9) Technical Rep~~ort~~

6. PERFORMING ORG. ~~REPORT NUMBER~~
186

7. AUTHOR(s)
10) J. S. Rustagi

8. ~~CONTRACT OR GRANT NUMBER(s)~~
15) N00014-78-C-0543

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Statistics
Ohio State University
1958 Neil Avenue, Columbus, Ohio 43210

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
NR 042-403

11. CONTROLLING OFFICE NAME AND ADDRESS
Office of Naval Research
Department of Navy
Arlington, Virginia 22207

12. ~~REPORT DATE~~
11) Aug~~ust 19~~79

13. ~~NUMBER OF PAGES~~
15

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
12) 17 p

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)
14) TR-186

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Variational methods, optimizing techniques, robustness, penalized maximum likelihood

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Recent applications of variational methods to problems of statistical inference are given. The problems of robustness, inadmissibility and penalized maximum likelihood estimates are discussed to exhibit their variational character.

406 331

Recent Applications of Variational Techniques in Statistics

1. Introduction

Statistical problems in general require extensive use of optimizing techniques. A variety of these methods, depending on the nature of the problem, have been used by statisticians. In recognition of the importance of optimizing methods in statistics, two research conferences in this area were recently held, and the proceedings of the conferences exhibit the wide variety of statistical applications in which an important part is played by optimization, Rustagi (1971, 1979).

Optimizing methods have been classified in four main categories: classical optimizing methods, mathematical programming methods, numerical methods and variational methods. A review of these methods was recently given in a paper by Rustagi (1978a). Some recent applications of optimization in statistics appear in a special issue of Communications in Statistics edited by Rustagi (1978a). Survey of some of the commonly used variational methods with their applications in statistics has also been given in a book, Rustagi (1976).

In this paper, some recent applications of variational techniques are given. Examples are provided from robustness studies, decision theory and estimation of probability densities.

Under variational methods we include all the techniques which are required to optimize a functional over a function space. In its simplest form, a variational problem results if one wants to optimize an integral of a known function of an unknown function and possibly of its derivative. In a sense variational methods correspond to methods of maxima and minima in calculus to similar methods in functional analysis.

Some of the early results in robustness studies required optimizing the variance of M-estimates introduced by Huber (1972) over the class of symmetric distributions. Such a criterion can also be stated in terms of minimizing Fisher's information and explicit solutions of these problems require variational methods. Recent extensions of the applications of variational techniques to dependent situations has been discussed by Portnoy (1977) and the method of geometry of moment spaces has been utilized by Collins and Portnoy (1979) for more general situations. This problem is discussed in section 2.

Questions of admissibility of certain decision problems arise in many contexts. Brown (1971) has recently studied the admissibility of certain decision functions in the multivariate-normal case under quadratic loss criterion. These questions lead naturally to the solutions of the variational problems. Using classical theory of calculus of variations and Euler-Lagrange equations, the inadmissibility of the decision function was exhibited with the help of nonexistence of the solution of an optimization problem. This novel application of a variational technique is especially illuminating as it provides a method of verifying admissibility under fairly general conditions. We discuss further details in section 3.

Estimation of densities utilizing penalized maximum likelihood methods, has been discussed by Good (1971), and Good and Gaskins (1971). A general formulation of the optimization problem arising from the above in abstract setting has been given by DeMontricher, Tapia and Thompson (1975). The existence of the solution of the proposed optimization problems has been proved and certain characterizations of the optimum solution are given. This topic is discussed in Section 4.

2. Robust statistics and variational techniques

There are several situations where variational techniques play an important role in the study of robust statistics. In this section we consider only two such examples. The first example is due to Huber (1972), Portnoy (1977) and Portnoy and Collins (1979), and the second is due to Bickel (1965).

In Huber's notation, we consider the M-estimates of a location parameter θ for the probability density $f(x-\theta)$, with c.d.f. $F(x-\theta)$ based on a random sample $X_1,X_2,\dots,X_n$. T_n is called an M-estimate for θ if it maximizes

$$\sum_{i=1}^{n} \rho(x_i - T_n) ,$$

where ρ is a given metric. M-estimates are also obtained if we solve the following type of equation

$$\sum_{i=1}^{n} \psi(x_i - T_n) = 0$$

where $\psi = \rho'$. M-estimates include least-squares and maximum likelihood estimates as examples, if we choose $\rho(x) = -x^2$ and $\rho(x) = -\frac{f'(x)}{f(x)}$ respectively.

Under fairly general conditions, it is known that

$$T_n = T(F_n) \rightarrow T(F) \text{ a.s. as } n \rightarrow \infty ,$$

such that

$$\int \psi(x-T(F))\, F(dx) = 0 .$$

Further, the asymptotic distribution of

$$\sqrt{n}(T_n - T(F))$$

is normal with mean 0 and variance $V(\psi,F)$ given by

$$V(\psi,F) = \int \left[\frac{\psi(x - T(F))}{\int \psi'(x - T(F))\, F(dx)} \right]^2 F(dx) = \frac{E(\psi^2)}{E(\psi'^2)} . \tag{2.1}$$

One of the important problems in robust estimation is to find the class of statistics $\{T_n\}$ such that asymptotic variance is minimized over the class $\mathcal{C}$ of all distribution functions F.

Robust estimates can also be generated by using linear combinations of order statistics and by the use of statistics derived from rank tests. These estimates are called L- and R-estimates, respectively. Consider T_n as a functional of the empirical distribution function F_n given by

$$T_n = T(F_n).$$

Then M-estimates are defined by the formula

$$\int \psi(x - T(F))\, F(dx) = 0 . \tag{2.2}$$

L-estimates are given by

$$T(F) = \int J(t)\, F^{-1}(t)dt \tag{2.3}$$

and R-estimates are defined by

$$\int J \left\{ \frac{1}{2}[F(x) + 1 - F(2T(F) - x)] \right\} F(dx) = 0$$

where the function J gives weights in linear combination of order statistics.

Restricting the estimates to the translation invariant class and assuming that the distribution function is symmetric, Huber (1964) showed that the distribution which minimizes Fisher information plays a major role in determining robust estimates.

Consider for example the class of all ϵ-contaminated normal distributions denoted by $\mathcal{C}$ where an element of the class $\mathcal{C}$ is denoted by

$$F(x) = (1 - \epsilon)\ \Phi(x) + \epsilon\ H(x),\ 0 \leq \epsilon < 1,$$

ϵ is known, Φ is the standard normal cumulative distribution function and $H(x)$ is a symmetric distribution function. Define $I(F) = \sup_{\psi}[V(\psi,F)]^{-1}$, with $\int \psi^2 dF \neq 0$. Then, Huber (1964) proved the following theorems.

Theorem 1. $I(F) < \infty$ if and only if F has absolutely continuous density $f(x)$ such that $\int (\frac{f'(x)}{f(x)})^2\ f(x)\ dx < \infty$, and then

$$I(F) = \int (\frac{f'(x)}{f(x)})^2 f(x)\ dx.$$

Theorem 2. If $\inf_{F \epsilon \mathcal{C}} I(F) = a < \infty$, then there exists a unique $F_0 \epsilon \mathcal{C}$ such that

$$I(F_0) = a.$$

Using variational techniques and some guesswork, F_0 and the corresponding ψ_0 can be found. For example in the above case we have the least favorable density $f_0(x)$ given by

$$f_0(x) = \frac{1-\epsilon}{\sqrt{2\pi}}\ e^{-\rho_0(x)}$$

$$\text{where } \rho_0(x) = \begin{cases} \frac{x^2}{2}, \text{ if } |x| < k \text{ ,} \\ k\ |x| - \frac{k^2}{2}, \ |x| \geq k \text{ .} \end{cases}$$

k is chosen so that

$$\frac{\varepsilon}{1-\varepsilon} = \frac{2}{k}\ \phi(k) - 2\Phi(-k) \text{ .}$$

Similar results hold for L- and R-estimates.

Robustness questions related to dependent situations, have been recently studied by Portnoy (1977) and Collins and Portnoy (1979). The variational problems arising in these applications require modern methods such as those of geometry of moment spaces, for reference, see Rustagi (1976). Consider the time-series moving average model,

$$X_i = \theta + Y_i + \rho\, Y_{i-1} + \rho\, Y_{i+1} \text{ ,}$$
$$i = 1,2,\ldots,n$$

where (i) $Y_0 = Y_n$, $Y_1 = Y_{n+1}$

(ii) θ = location parameter

(iii) $X_1,\ldots,X_n$ have a stationary distribution

(iv) $|\rho| < 1$

(v) $Y_1, Y_2, \ldots, Y_n$ are independently and identically distributed random variables having continuous and symmetric c.d.f. G with p.d.f. g(y).

The approximate asymptotic variance of the estimates of θ is given by

$$V_1(g,\psi) = \frac{E(\psi^2(Y))}{[E(\psi'(Y))]^2} + 4\rho\, \frac{E(Y\psi(Y))}{E(\psi'(Y))} + O(\rho^2).$$

Variational methods are used to find ϕ_1 which minimizes the variance for a fixed g and turns out to be the same as in the independent case.

The second example in the study of robust estimates of location is due to Bickel (1965). Variational problems occur naturally in finding minimum efficiency. In this paper, Bickel considers minimum efficiency with respect to the class of all symmetric and symmetric unimodal distributions, of the Winsorized and trimmed means with respect to the mean.

Suppose $W_1 < W_2 < \ldots < W_n$ are the ordered statistic of a sample $X_1,\ldots,X_n$ from an absolutely continuous distribution function $F(x)$. Then α-trimmed mean is defined by

$$\bar{X}_\alpha = \frac{\sum_{i=[\alpha n+1]}^{n-[\alpha n]} W_i}{n - 2[\alpha n]}$$

where $[\alpha n]$ is the greatest integer in αn, $0 \leq \alpha < \frac{1}{2}$.

α-Winsorized mean is defined by

$$X_\alpha^* = \frac{1}{n}\left\{[\alpha n]\, W_{[\alpha n]} + \sum_{i=[\alpha n]+1}^{n-[\alpha n]} W_i + [\alpha n]\, W_{n-[\alpha n]+1}\right\}$$

Let $e_1(\alpha)$ and $e_2(\alpha)$ be the asymptotic relative efficiencies of $\bar{X}_\alpha$ and X_α^* with respect to $\bar{X}$ respectively. Then

$$e_1(\alpha) = (1-2\alpha)^2 \left(\int_{-\infty}^{\infty} x^2 f(x)dx\right) \left(\int_{-\lambda}^{\lambda} x^2 f(x)dx \right.$$

$$\left. + 2\, \alpha x(\alpha)^2\right)^{-1}$$

and

$$e_2(\alpha) = \frac{\int_{-\infty}^{\infty} x^2 f(x)dx}{c + 2\alpha(\lambda + \frac{\alpha}{k})^2}$$

where

(1) $x(\alpha) = -\lambda$

(2) $f(x(\alpha)) = k$

(3) $\int_{-\lambda}^{\lambda} x^2 f(x)dx = c$

Let $\mathcal{F}$ be the class of all symmetric unimodal distribution functions. Then the following theorem is proved.

Theorem:
$$\inf_{F \in \mathcal{F}} e_1(\alpha) = \frac{1}{1+4\alpha}$$
$$\inf_{F \in \mathcal{F}} e_2(\alpha) = \frac{1}{3} .$$

Using Lagrange's method of undetermined multiples, Euler-Lagrange equations of calculus of variations provide the minimizing densities $f(x)$. Detailed proofs are in Bickel (1965).

3. Admissibility questions and variational methods

Necessary and sufficient conditions for an estimator of the mean of a multivariate normal distribution under squared loss function, to be admissible have been discissed by Brown (1971). The problem of admissibility is directly related to problems of diffusion. This correspondence is established through classical variational techniques using Euler-Lagrange equation.

Let $p_\theta(x)$ be the m-dimensional multivariate normal density of the random vector X. Let $\delta(x)$ denote an estimate of θ. Suppose the loss function is given by

$$L(\theta,\delta) = (\delta - \theta)'D(\delta - \theta)$$

where D is known diagonal matrix. We have the following notation:

$R(\theta,\delta) = E_\theta[L(\theta,\delta(x))]$

$G(\theta)$ = Prior distribution function of θ.

$B(G,\delta) = \int R(\theta,\delta)\, G(d\theta)$,

= Bayes risk.

It is well-known that the Bayes estimator is given by

$$\delta_G(X) = \frac{\int \theta p_\theta(x)\, G(d\theta)}{\int p_\theta(x)\, G(d\theta)} \quad . \tag{3.1}$$

Let $||y||^2 = y'Dy$. When $D = I$, $||y||^2 = |y|^2 = \sum_{i=1}^{m} y_i^2$. Suppose

$$g^*(x) = \int p_\theta(x)\, G(d\theta)$$

and

$$\nabla g^*(x) = \left(\frac{\partial g^*}{\partial X_1}, \frac{\partial g^*}{\partial X_2}, \ldots, \frac{\partial g^*}{\partial X_m}\right)' \quad .$$

Then

$$\delta_G(x) = x + \frac{\nabla g^*(x)}{g^*(x)} \tag{3.2}$$

Define

$$\gamma_G(x) = \delta_G(x) - x \quad . \tag{3.3}$$

The necessary and sufficient conditions for admissibility were given by Stein (1955). One of the conditions for an estimator $\delta_F(x)$ with prior F to be admissible is given by the following sufficient condition.

Stein's condition: δ_F is admissible only if there exist non-negative finite Borel measures G_i, $i = 1,2,\ldots$ with G_i having compact support with $G_i(\{0\}) = 1$ such that

$$B(G_i,\delta_F) - B(G_i,\delta_{G_i}) \to 0 \tag{3.4}$$

as $i \to \infty$.

Condition (3.4) can be written in the following form if we define f and f* related to c.d.f. F, in the same way as we defined g and g* related to c.d.f. G.

$$B(G_i,\delta_F) - B(G_i,\delta_{G_i}) = \int \left|\left|\frac{\nabla f^*(x)}{f^*(x)} - \frac{\nabla g^*(x)}{g^*(x)}\right|\right|^2 g_i^*(x)dx$$

$$= \int ||\nabla j_i(x)||^2 f^*(x)dx = I(j_i(x))$$

where

$$j_i(x) = \left(\frac{g_i^*(x)}{f_i^*(x)}\right)^{\frac{1}{2}} .$$

The condition of admissibility is then reduced to the problem of minimizing

$$I(j(x)) \tag{3.5}$$

subject to the constraints

(i) $j(x) \geq 1, \ |x| \leq 1$

(ii) $\lim_{\gamma \to \infty} \ \sup_{|x|=r} \ j(x) = 0$.

Euler-Lagrange equation for (3.5) is given by

$$f^*(x) \sum_{i=1}^{m} \frac{\partial^2 j(x)}{\partial x_i^2} + \sum_{i=1}^{m} \frac{\partial f^*(x)}{\partial x_i} \frac{\partial j(x)}{\partial x_i} = 0$$

This is an elliptic partial differential equation and its solution provides an answer to the admissibility question posed above. Brown has used elaborate machinery to show that the solution to the elliptic differential equation exists for $|x| > 1$ which shows the inadmissibility of the usual Bayes estimator of the multivariate normal mean θ for $m > 2$.

4. Variational methods and penalized maximum likelihood estimates

A method of estimating probability densities utilizing penalty functions was introduced by Good (1971) and was developed further by Good and Gaskins (1971). To remove roughness in estimating the probability density functions, Good and Gaskins require maximizing not the log likelihood but maximizing log likelihood adjusted by a known function of the density function. The optimization problems so introduced lead naturally to variational problems. Many such problems in their abstract form have recently been studied by DeMontricher, Tapia and Thompson (1975).

Given that $X_1, X_2, \ldots, X_n$ is a random sample of size n from an unknown density function f(x), the penalized maximum likelihood estimates of f are defined by maximizing

$$L(f) = \prod_{i=1}^{n} f(x_i)\, e^{-\Phi(f)} \tag{4.1}$$

subject to the constraints

$$\int f(x)dx = 1 \tag{4.2}$$

and

$$f(x) \geq 0 \; . \tag{4.3}$$

In a little more abstract form, the problem is formulated in terms of the following notation. Let

Ω = interval (a, b) ,

$L'(\Omega)$ = class of Lebesgue integrable functions and $f \in L'(\Omega)$

$H(\Omega)$ = manifold in $L'(\Omega)$

$\Phi: \; H(\Omega) \rightarrow R$.

The variational problem is maximizing (4.1) over a class $H(\Omega)$ subject to constraints (4.2) and (4.3) for all $x \in \Omega$. The existence and uniqueness of the maximizing f is given by the following theorem due to DeMontricher, Tapia and Thompson.

Theorem 1. Suppose $H(\Omega)$ is a reproducing Kernel Hilbert space, and integration over Ω is a continuous functional and there exists at least one f with

$$f(x) \geq 0, \int f dx = 1 \text{ and } f(x_i) > 0 \text{ ,}$$
$$i = 1,2,\dots,n \text{ for all } x \in \Omega.$$

Then the maximum penalized estimate corresponding to $H(\Omega)$ exists and is unique.

Under certain additional assumptions, the solution of the above problem can be characterized as a polynomial spline. Motivated by information theoretic considerations, Good and Gaskins considered the first penalized maximum likelihood estimate of the density function by using

$$\Phi(f) = \alpha \int_{-\infty}^{\infty} \frac{f'(t)^2}{f(t)} dt, \ \alpha > 0$$
$$= 4\alpha \int \frac{(d\sqrt{f})^2}{dt} \ dt \ . \qquad (4.4)$$

Assume that $H(\Omega)$ is such that

$$\sqrt{f} \in H'(-\infty,\infty) \ .$$

The functional to be optimized is still

$$\prod_{i=1}^{n} f(x_i)\ e^{-\Phi(f)} \qquad (4.5)$$

Suppose $u = \sqrt{f}$, then the optimization problem above is of the following form

$$\text{Max} \prod_{i=1}^{n} u^2(x_i)\, e^{-4\alpha \int u'(t)^2 dt}$$

subject to the constraints

$$u \in H'(-\infty,\infty)$$
$$\text{and} \int u^2(t)dt = 1 . \qquad (4.6)$$

The authors show that the first maximum likelihood penalized estimate of Good and Gaskins exists and is unique.

The second maximum likelihood penalized estimator is defined with help of

$$\Phi(f) = \alpha \int_{-\infty}^{\infty} f'(t)^2 dt + \beta \int f''^2(t)dt \qquad (4.7)$$

for some $\alpha \geq 0$ and $\beta > 0$.

Although in this case also, one can show that the estimate exists and is unique, it is not possible to obtain the estimate by an approach provided by Good and Gaskins.

5. Comments

The wide variety of applications of variational techniques exemplified above by various examples, exhibits their importance as a necessary tool for a statistician. Once the problem can be formulated in the form in which its variational character is apparent, there are many available techniques to solve it. There are, however, a large class of problems which need further study. Consider the problem of feedback control where the equations governing the motion of a particle are not known. Suppose these equations are estimated from data. The dynamic programming solution to such a feedback problem requires a different approach and the statistical dynamic programming solution then naturally leads to open questions. Distributions and stochastic convergence of the solution are now needed and interpretation of the optimal policy is required in view of the estimated relations.

6. References

1. Bickel, Peter (1965). On some robust estimates of location, Ann. Math. Statist. 36, 847-858.

2. Brown, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems, Ann. Math. Statist. 42, 885-903.

3. De Montricher, G. F., Tapia, R. A., and Thompson, J. F. (1975). Nonparametric maximum likelihood estimation of probability densities by penalty function methods, Ann. Statist. 3, 1329-1348.

4. Good, I. J. (1971). A nonparametric roughness penalty for probability densities, Nature 229, 29-30.

5. Good, I. J. and Gaskins, A. (1971). Nonparametric roughness penalties for probability densities, Biometrika 58, 255-277.

6. Huber, P. (1964). Robust estimation of a location parameter, Ann. Math. Statist. 35, 73-101.

7. Huber, P. (1972). Robust Statistics - A review, Ann. Math. Statist. 43, 1042-1067.

8. Portnoy, Stephen L. (1977). Robust Estimates in dependent situations, Ann. Statist. 5, 22-43.

9. Portnoy, Stephen L. and Collins, J. R. (1979). Maximizing the variance of M-estimators using the generalized method of moment spaces, Technical Report, Department of Mathematics, University of Alberta, Edmonton, Alberta.

10. Rustagi, J. S. (Editor): Optimizing Methods in Statistics, Academic Press, New York, 1971.

11. Rustagi, J. S.: Variational Methods in Statistics, Academic Press, New York, 1976.

12. Rustagi, J. S. (1978). Optimization in Statistics, Comm. Statist.-Simula. Computa. B7(4), 303-307.

13. Rustagi, J. S. (Editor) (1978). Special Issue, Comm. Statist.-Simula. Computa. B7(4), 303-435.

14. Rustagi, J. S. (Editor) (1979). Optimizing Methods in Statistics, Academic Press, New York.

15. Stein, C. (1955). A necessary and sufficient condition for admissibility. Ann. Math. Statist. 26, 518-522.